

Relative α -Entropy Minimizers Subject to Linear Statistical Constraints

M. Ashok Kumar
Department of ECE
Indian Institute of Science
Bangalore, Karnataka 560012, India
Email: ashokm@ece.iisc.ernet.in

Rajesh Sundaresan
Department of ECE
Indian Institute of Science
Bangalore, Karnataka 560012, India
Email: rajeshs@ece.iisc.ernet.in

Abstract—We study minimization of a parametric family of relative entropies, termed relative α -entropies (denoted $\mathcal{J}_\alpha(P, Q)$). These arise as redundancies under mismatched compression when cumulants of compressed lengths are considered instead of expected compressed lengths. These parametric relative entropies are a generalization of the usual relative entropy (Kullback-Leibler divergence). Just like relative entropy, these relative α -entropies behave like squared Euclidean distance and satisfy the Pythagorean property. Minimization of $\mathcal{J}_\alpha(P, Q)$ over the first argument on a set of probability distributions that constitutes a linear family is studied. Such a minimization generalizes the maximum Rényi or Tsallis entropy principle. The minimizing probability distribution (termed \mathcal{J}_α -projection) for a linear family is shown to have a power-law.

I. INTRODUCTION

Maximum entropy principle is a well-known selection principle under uncertainty. This is an idea that dates back to L. Boltzmann, was popularized by E. T. Jaynes [1], and has its foundation in the theory of large deviation. Suppose that an ensemble average measurement (say sample mean, sample second moment, or any other similar linear statistic) is made on the realization of a sequence of iid random variables. The realization must then have an empirical distribution that obeys the constraint placed by the measurement – the empirical distribution must belong to an appropriate convex set, say \mathbb{E} . Large deviation theory tells us that a special member of \mathbb{E} , denoted P^* , is overwhelmingly more likely than the others. If the alphabet \mathbb{X} is finite (with cardinality $|\mathbb{X}|$), and the prior probability distribution (before measurement) is the uniform distribution U on \mathbb{X} , then P^* is the one that minimizes the relative entropy¹

$$\mathcal{J}(P||U) = \log |\mathbb{X}| - H(P),$$

¹The relative entropy of P with respect to Q is defined as

$$\mathcal{J}(P||Q) := \sum_{x \in \mathbb{X}} P(x) \log \frac{P(x)}{Q(x)}$$

and the Shannon entropy of P is defined as

$$H(P) := - \sum_{x \in \mathbb{X}} P(x) \log P(x).$$

The usual convention is $p \log \frac{p}{q} = 0$ if $p = 0$ and $+\infty$ if $p > q = 0$.

which is the same as the one that maximizes (Shannon) entropy², subject to $P \in \mathbb{E}$. In Jaynes' words, "... it is maximally noncommittal to the missing information" [1].

As a physical example, let us tag a particular molecule in the atmosphere. Let X denote the height of the molecule in the atmosphere. Then the potential energy of the molecule is mgX . Let us suppose that the average potential energy is held constant, that is, $E[mgX] = c$, a constant. Then the probability distribution of the height of the molecule is taken to be the exponential distribution $\lambda \exp(-\lambda x)$, where $\lambda = mg/c$. This is also the maximum entropy probability distribution subject to first moment constraint [2].

More generally, if the prior probability distribution (before measurement) is Q , then P^* minimizes $\mathcal{J}(P||Q)$ subject to $P \in \mathbb{E}$. Something more specific can be said: P^* is the limiting conditional probability distribution of a "tagged" particle under the conditioning imposed by the measurement. This is called *the conditional limit theorem* or the *Gibbs conditioning principle*; see for example Campenhout and Cover [3] or Csiszár [4] for a more general result.

It is well-known that $\mathcal{J}(P||Q)$ behaves like "squared Euclidean distance" and has the "Pythagorean property" (Csiszár [5]). In view of this and since P^* minimizes $\mathcal{J}(P||Q)$ subject to $P \in \mathbb{E}$, one says that P^* is "closest" to Q in the relative entropy sense amongst the probability distributions in \mathbb{E} , or in other words, " P^* is the \mathcal{J} -projection of Q on \mathbb{E} ". Motivated by the above maximum entropy and Gibbs conditioning principles, \mathcal{J} -projection was extensively studied by Csiszár [4], [5], and Csiszár and Matúš [6].

This paper is on the projection problem associated with a parametric generalization of relative entropy. To see how this parametric generalization arises, we return to our remark on how relative entropy arises in Shannon theory. For this, we must first recall how Rényi entropies are a parametric generalization of the Shannon entropy.

Rényi entropies $H_\alpha(P)$ for $\alpha \in (0, 1)$ play the role of Shannon entropy when the *normalized cumulant* of compression length,

$$\frac{1}{\rho} \log E[\exp\{\rho L(X)\}],$$

²Hence the name maximum entropy principle.

is considered instead of expected compression length $E[L(X)]$, where $\rho > 0$ is the cumulant parameter. Campbell [7] showed that the minimum normalized cumulant subject to all compression strategies that satisfy the Kraft inequality is

$$H_\alpha(P) = \frac{1}{1-\alpha} \log \sum_{x \in \mathbb{X}} P(x)^\alpha,$$

where $\alpha = 1/(1+\rho)$. We also have $\lim_{\alpha \rightarrow 1} H_\alpha(P) = H(P)$, so that Rényi entropy may be viewed as a generalization of Shannon entropy.

If the compressor assumed that the true probability distribution is Q , instead of P , then the gap in the normalized cumulant's optimal value is an analogous parametric divergence quantity³, which we shall denote $\mathcal{J}_\alpha(P, Q)$ [9]. The same quantity also arises when we study the gap from optimality of mismatched guessing exponents. See Arikan [10] and Hanawal and Sundaresan [11] for general results on guessing, and see Sundaresan [12] and [9] on how $\mathcal{J}_\alpha(P, Q)$ arises in the context of mismatched guessing. Recently, Bunte and Lapidoth [13] have shown that the $\mathcal{J}_\alpha(P, Q)$ also arises as redundancy in a mismatched version of the problem of coding for tasks.

\mathcal{J}_α may be expressed as

$$\begin{aligned} \mathcal{J}_\alpha(P, Q) &= \frac{\alpha}{1-\alpha} \log \left[\sum_x P(x) Q(x)^{\alpha-1} \right] - \frac{1}{1-\alpha} \log \sum_x P(x)^\alpha \\ &\quad + \log \sum_x Q(x)^\alpha. \end{aligned} \quad (1)$$

For each $\alpha > 0, \alpha \neq 1$, $\mathcal{J}_\alpha(P, Q) \geq 0$ with equality iff $P = Q$ [9, Prop. 4]. Also, $\mathcal{J}_\alpha(P, Q) = \infty$ only when either

- $\alpha < 1$ and $\text{Supp}(P) \not\subseteq \text{Supp}(Q)$ ⁴ or
- $\alpha > 1$ and $\text{Supp}(P) \cap \text{Supp}(Q) = \emptyset$.

For $\alpha > 1$, $\mathcal{J}_\alpha(P, Q)$ turns out to be relevant in a robust parameter estimation problem of statistics [14].

As one might expect, it is known that (see for example, Sundaresan [9, Sec. V-5]) or Johnson and Vignat [15, A.1]) $\lim_{\alpha \rightarrow 1} \mathcal{J}_\alpha(P, Q) = \mathcal{J}(P||Q)$, so that we may think of relative entropy as $\mathcal{J}_1(P, Q)$. Thus \mathcal{J}_α is a generalization of relative entropy, i.e., a *relative α -entropy*⁵.

Not surprisingly, the maximum Rényi entropy principle has been considered as a natural alternative to the maximum entropy principle of decision making under uncertainty. This principle is equivalent to another principle of maximizing the so-called Tsallis entropy which happens to be a monotone function of the Rényi entropy. Rényi entropy maximizers under moment constraints are distributions with a power-law decay (when $\alpha < 1$). See Costa et al. [17] or Johnson and Vignat [15]. Many statistical physicists have studied this principle in the hope that it may “explain” the emergence of power-laws

in many naturally occurring physical and socio-economic systems, beginning with Tsallis [18]. Based on our explorations of the vast literature on this topic, we feel that our understanding, particularly one that ought to involve a modeling of the *dynamics* of such systems with the observed power-law profiles as equilibria in the asymptotics of large time, is not yet as mature as our understanding of the classical Boltzmann-Gibbs setting. But, by noting that $\mathcal{J}_\alpha(P, U) = \log |\mathbb{X}| - H_\alpha(P)$, we see that both the maximum Rényi entropy principle and the maximum Tsallis entropy principle are particular instances of a “minimum relative α -entropy principle”:

$$\text{minimize } \mathcal{J}_\alpha(P, Q) \text{ over } P \in \mathbb{E}.$$

We shall call the minimizing P^* as the \mathcal{J}_α -projection of Q on \mathbb{E} .

The objective of this paper is to characterize the \mathcal{J}_α -projection on a particular convex family called *linear family* which we shall define now.

Definition 1: A linear family characterized by k functions $f_i : \mathbb{X} \rightarrow \mathbb{R}$, $1 \leq i \leq k$, is the set of probability distributions given by

$$\mathbb{L} = \left\{ P \in \mathcal{P}(\mathbb{X}) : \sum_x P(x) f_i(x) = 0, i = 1, \dots, k \right\}. \quad (2)$$

We will show that the \mathcal{J}_α -projection on a linear family comes from an α -power law family analogous to the fact that the \mathcal{J}_1 -projection on such a family comes from an exponential family [19].

II. CHARACTERIZATION OF \mathcal{J}_α -PROJECTION ON A LINEAR FAMILY

In this section, we find the structure of the \mathcal{J}_α -projection on a linear family \mathbb{L} and prove a necessary and sufficient condition for a $P^* \in \mathbb{L}$ to be the \mathcal{J}_α -projection on \mathbb{L} . We consider the cases $\alpha > 1$ and $\alpha < 1$ separately in the two subsections. The proof for the $\alpha < 1$ case is similar to Csiszár and Shields' proof for $\alpha = 1$ case [19]. For the proof of $\alpha > 1$ case, we will resort to the Lagrange multiplier technique.

A. $\alpha < 1$:

The result for $\alpha < 1$ is the following.

Theorem 2: Let $\alpha < 1$. Let \mathbb{L} be a linear family characterized by $f_i, i = 1, \dots, k$. Let Q be a probability distribution with $\text{Supp}(Q) = \mathbb{X}$. Then the following hold.

- Q has an \mathcal{J}_α -projection on \mathbb{L} . Call it P^* .
- $\text{Supp}(P^*) = \text{Supp}(\mathbb{L})$ ⁶ and the Pythagorean equality holds (see Figure 1):

$$\mathcal{J}_\alpha(P, Q) = \mathcal{J}_\alpha(P, P^*) + \mathcal{J}_\alpha(P^*, Q) \quad \forall P \in \mathbb{L}. \quad (3)$$

⁶ $\text{Supp}(\mathbb{L})$ is defined to be the union of supports of members of \mathbb{L} .

³Blumer and McEliece [8], in their attempt to find better upper and lower bounds on the redundancy of generalized Huffman coding, were indirectly bounding this parameterized divergence.

⁴ $\text{Supp}(P) = \{x : P(x) > 0\}$.

⁵This terminology is from Lutwak, et al. [16].

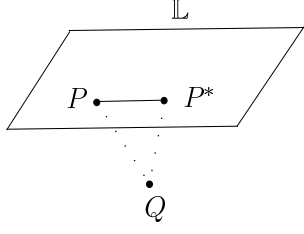


Fig. 1. Pythagorean property

(c) The \mathcal{J}_α -projection P^* satisfies

$$P^*(x) = Z^{-1} \cdot \left[Q(x)^{\alpha-1} + (1-\alpha) \sum_{i=1}^k \theta_i^* f_i(x) \right]^{\frac{1}{\alpha-1}} \quad \forall x \in \mathbb{X}, \quad (4)$$

where $\theta_1^*, \dots, \theta_k^*$ are scalars and Z is the normalization constant that makes P^* a probability distribution.

(d) The \mathcal{J}_α -projection is unique. \square

Proof: (a) From (1), it is clear that the mapping $P \mapsto \mathcal{J}_\alpha(P, Q)$ is continuous as each of the terms in (1) is continuous in P . Also \mathbb{L} is compact. Hence the \mathcal{J}_α -projection exists.

(b) Let $P \in \mathbb{L}$ and let $P_t = (1-t)P^* + tP$, $0 \leq t \leq 1$. Since $P_t \in \mathbb{L}$, the mean value theorem says that for each $t \in (0, 1)$, there exists $\tilde{t} \in (0, t)$ such that

$$0 \leq \frac{1}{t} [\mathcal{J}_\alpha(P_t, Q) - \mathcal{J}_\alpha(P^*, Q)] = \frac{d}{ds} \mathcal{J}_\alpha(P_s, Q) \big|_{s=\tilde{t}}. \quad (5)$$

The first inequality follows from the fact that P^* is the projection. Using (1), we see that

$$\frac{d}{ds} \mathcal{J}_\alpha(P_s, Q) = \frac{\alpha}{1-\alpha} \left[\frac{\sum_x (P(x) - P^*(x)) Q(x)^{\alpha-1}}{\sum_x P_s(x) Q(x)^{\alpha-1}} - \frac{\sum_x (P(x) - P^*(x)) P_s(x)^{\alpha-1}}{\sum_x P_s(x)^\alpha} \right]. \quad (6)$$

As $t \downarrow 0$, from (6) and the inequality in (5), we get

$$\begin{aligned} & \frac{\sum_x (P(x) - P^*(x)) Q(x)^{\alpha-1}}{\sum_x P^*(x) Q(x)^{\alpha-1}} \\ & \geq \frac{\sum_x (P(x) - P^*(x)) P^*(x)^{\alpha-1}}{\sum_x P^*(x)^\alpha}. \end{aligned} \quad (7)$$

That is,

$$\frac{\sum_x P(x) Q(x)^{\alpha-1}}{\sum_x P^*(x) Q(x)^{\alpha-1}} \geq \frac{\sum_x P(x) P^*(x)^{\alpha-1}}{\sum_x P^*(x)^\alpha},$$

which, using (1), can be seen to be equivalent to (3), but with inequality “ \geq ” in place of equality.

Suppose $P^*(x) = 0$ for an $x \in \text{Supp}(P)$. Then $\alpha < 1$ implies that right-hand side of (6) goes to $-\infty$ as $t \downarrow 0$, which contradicts the nonnegativity requirement in (5). Hence $\text{Supp}(P) \subseteq \text{Supp}(P^*)$. Since P was arbitrary, we have $\text{Supp}(P^*) = \text{Supp}(\mathbb{L})$.

We will now establish equality in (3). Once again let $P \in \mathbb{L}$. Since $\text{Supp}(P^*) = \text{Supp}(\mathbb{L})$, we can find a new $\tilde{t} < 0$ such that $P_t = (1-t)P^* + tP$ is a valid probability distribution for all $t \in (\tilde{t}, 0)$. Hence $P_t \in \mathbb{L}$ for all $t \in (\tilde{t}, 0)$. Since $t < 0$, we have

$$0 \geq \frac{1}{t} [\mathcal{J}_\alpha(P_t, Q) - \mathcal{J}_\alpha(P^*, Q)] \quad \forall t \in (\tilde{t}, 0).$$

An argument similar to the one that led to (7) now proves that (7) holds with “ \leq ” as well. This proves (3) and completes the proof of (b).

(c) This follows the proof of Csiszár and Shields for case $\alpha = 1$ [19, Th. 3.4].

From (2), it is clear that the probability distributions $P \in \mathbb{L}$, when considered as $|\text{Supp}(\mathbb{L})|$ -dimensional vectors, belong to the orthogonal complement \mathcal{F}^\perp of the subspace \mathcal{F} of $\mathbb{R}^{|\text{Supp}(\mathbb{L})|}$ spanned by the vectors $f_i(\cdot)$, $i = 1, \dots, k$, restricted to $\text{Supp}(\mathbb{L})$. These $P \in \mathbb{L}$ actually span \mathcal{F}^\perp . (This follows from the fact that if a subspace of $\mathbb{R}^{|\text{Supp}(\mathbb{L})|}$ contains a vector all of whose components are strictly positive, here P^* , then it is spanned by the probability vectors of that space.) Using (3), which is the same as (7) with equality, we have

$$\sum_x P(x) \left(\frac{Q(x)^{\alpha-1}}{\sum_a P^*(a) Q(a)^{\alpha-1}} - \frac{P^*(x)^{\alpha-1}}{\sum_a P^*(a)^\alpha} \right) = 0 \quad \forall P \in \mathbb{L}.$$

Consequently, the vector

$$\frac{Q(\cdot)^{\alpha-1}}{\sum_a P^*(a) Q(a)^{\alpha-1}} - \frac{P^*(\cdot)^{\alpha-1}}{\sum_a P^*(a)^\alpha}$$

belongs to $(\mathcal{F}^\perp)^\perp = \mathcal{F}$, that is,

$$\frac{Q(x)^{\alpha-1}}{\sum_a P^*(a) Q(a)^{\alpha-1}} - \frac{P^*(x)^{\alpha-1}}{\sum_a P^*(a)^\alpha} = \sum_{i=1}^k \lambda_i f_i(x) \quad \forall x \in \text{Supp}(\mathbb{L}),$$

for some scalars λ_i , $i = 1, \dots, k$. This verifies (4) for obvious choices of Z and θ_i^* , $i = 1, \dots, k$.

(d) Let P_1^* and P_2^* be two projections of Q on \mathbb{L} . Then $\mathcal{J}_\alpha(P_1^*, Q) = \mathcal{J}_\alpha(P_2^*, Q)$. By (c), we have

$$\mathcal{J}_\alpha(P_2^*, Q) = \mathcal{J}_\alpha(P_2^*, P_1^*) + \mathcal{J}_\alpha(P_1^*, Q).$$

Canceling $\mathcal{J}_\alpha(P_2^*, Q)$ and $\mathcal{J}_\alpha(P_1^*, Q)$, we get $\mathcal{J}_\alpha(P_2^*, P_1^*) = 0$ which implies $P_1^* = P_2^*$. \blacksquare

One can also have a converse.

Theorem 3: Let $\alpha < 1$. Let $P^* \in \mathbb{L}$ be a probability distribution of the form (4). Then P^* satisfies (3) and P^* is the \mathcal{J}_α -projection Q on \mathbb{L} . \square

Proof: If $P^* \in \mathbb{L}$ is of the stated form, then since every $P \in \mathbb{L}$ satisfies

$$\sum_{i=1}^k P(x) f_i(x) = \sum_{i=1}^k P^*(x) f_i(x) = 0, \quad i = 1, \dots, k,$$

we have from (4) that

$$Z^{\alpha-1} \sum_x P(x) P^*(x)^{\alpha-1} = \sum_x P(x) Q(x)^{\alpha-1},$$

and

$$Z^{\alpha-1} \sum_x P^*(x)^\alpha = \sum_x P^*(x) Q(x)^{\alpha-1}.$$

Combining the above two equations to eliminate $Z^{\alpha-1}$, we get

$$\sum_x P(x) Q(x)^{\alpha-1} = \frac{\sum_x P^*(x) Q(x)^{\alpha-1}}{\sum_x P^*(x)^\alpha} \sum_x P(x) P^*(x)^{\alpha-1},$$

which, using (1), can be seen to be equivalent to (3). Thus, for any $P \in \mathbb{L}$, we have

$$\begin{aligned} \mathcal{J}_\alpha(P, Q) &= \mathcal{J}_\alpha(P, P^*) + \mathcal{J}_\alpha(P^*, Q) \\ &\geq \mathcal{J}_\alpha(P^*, Q), \end{aligned}$$

which implies that P^* is the \mathcal{J}_α -projection of Q on \mathbb{L} . ■

When $\alpha > 1$, in general, $\text{Supp}(P^*) \neq \text{Supp}(\mathbb{L})$ as shown by the following counterexample.

Take $\alpha = 2$ and $\mathbb{X} = \{1, 2, 3, 4\}$. Take $Q = (1/4, 1/4, 1/4, 1/4)$. Consider the linear family,

$$\mathbb{L} = \{P \in \mathcal{P}(\mathbb{X}) : 8p_1 + 4p_2 + 2p_3 + p_4 = 7\}.$$

Thus

$$\mathbb{L} = \left\{ P \in \mathcal{P}(\mathbb{X}) : \sum_x P(x) f_1(x) = 0 \right\},$$

where $f_1(\cdot) = (1, -3, -5, -6)$.

We claim that the \mathcal{J}_α -projection of Q on \mathbb{L} is $P^* = (3/4, 1/4, 0, 0)$. To check this claim, first note that $P^* \in \mathbb{L}$. Also, with $Z = 2/5$ and $\theta_1^* = -1/20$, we can check that

$$\begin{aligned} 0 &< ZP^*(x) = Q(x) + \theta_1^* f_1(x), \quad x = 1, 2, \\ 0 &= ZP^*(3) = Q(3) + \theta_1^* f_1(3), \\ 0 &= ZP^*(4) > Q(4) + \theta_1^* f_1(4). \end{aligned}$$

Hence, for any $P \in \mathbb{L}$,

$$\begin{aligned} Z^{\alpha-1} \sum_{x=1}^4 P(x) P^*(x)^{\alpha-1} &= Z \sum_{x=1}^4 P(x) P^*(x) \\ &\geq \sum_{x=1}^4 P(x) [Q(x) + \theta_1^* f_1(x)] \\ &= \sum_{x=1}^4 P(x) Q(x) \\ &= \sum_{x=1}^4 P(x) Q(x)^{\alpha-1}, \end{aligned} \quad (8)$$

where the penultimate equality follows because $P \in \mathbb{L}$. Similarly, one can show that

$$Z^{\alpha-1} \sum_x P^*(x)^\alpha = \sum_x P^*(x) Q(x)^{\alpha-1}.$$

Combining this with (8) to eliminate $Z^{\alpha-1}$, we get

$$\sum_x P(x) Q(x)^{\alpha-1} \leq \frac{\sum_x P^*(x) Q(x)^{\alpha-1}}{\sum_x P^*(x)^\alpha} \sum_x P(x) P^*(x)^{\alpha-1},$$

which, using (1), can be seen to be equivalent to

$$\begin{aligned} \mathcal{J}_\alpha(P, Q) &\geq \mathcal{J}_\alpha(P, P^*) + \mathcal{J}_\alpha(P^*, Q) \\ &\geq \mathcal{J}_\alpha(P^*, Q). \end{aligned} \quad (9)$$

Thus, P^* is the \mathcal{J}_α -projection of Q on \mathbb{L} .

Clearly $\text{Supp}(P^*) \subsetneq \text{Supp}(\mathbb{L})$. Also for $P = (0.8227, 0.0625, 0.0536, 0.0612) \in \mathbb{L}$, numerical calculations yield a strict inequality in (9) since the left-hand side and the right-hand side of (9) evaluate to 1.0114 and 0.9871, respectively.

As a consequence of this, the proof of Theorem 2 for the case $\alpha < 1$ cannot be carried forward to establish the structure of \mathcal{J}_α -projection for the case $\alpha > 1$. We will resort to the usual Lagrange multiplier technique for this case.

B. $\alpha > 1$:

We now establish the form of the \mathcal{J}_α -projection on a linear family when $\alpha > 1$.

Theorem 4: Let $\alpha > 1$. Let \mathbb{L} be a linear family characterized by $f_i, i = 1, \dots, k$. Let Q be a probability distribution with $\text{Supp}(Q) = \mathbb{X}$. Then the following hold.

- (a) Q has an \mathcal{J}_α -projection on \mathbb{L} . Call it P^* .
- (b) The \mathcal{J}_α -projection P^* satisfies

$$P^*(x) = Z^{-1} \cdot \left[Q(x)^{\alpha-1} + (1-\alpha) \sum_{i=1}^k \theta_i^* f_i(x) \right]_+^{\frac{1}{\alpha-1}} \quad \forall x \in \mathbb{X}, \quad (10)$$

where $\theta_1^*, \dots, \theta_k^*$ are scalars and Z is the normalization constant that makes P^* a probability distribution and $[u]_+ = \max\{u, 0\}$.

- (c) The Pythagorean inequality holds:

$$\mathcal{J}_\alpha(P, Q) \geq \mathcal{J}_\alpha(P, P^*) + \mathcal{J}_\alpha(P^*, Q) \quad \forall P \in \mathbb{L}. \quad (11)$$

- (d) The \mathcal{J}_α -projection is unique.
- (e) If $\text{Supp}(P^*) = \text{Supp}(\mathbb{L})$, then (11) holds with equality. □

Proof: (a) Same as proof of Theorem 2-(a).

(b) The optimization problem for the \mathcal{J}_α -projection is

$$\min_P \mathcal{J}_\alpha(P, Q) \quad (12)$$

$$\text{subject to } \sum_x P(x) f_i(x) = 0, \quad i = 1, \dots, k \quad (13)$$

$$\sum_x P(x) = 1 \quad (14)$$

$$P(x) \geq 0 \quad \forall x \in \mathbb{X}. \quad (15)$$

We will proceed in a sequence of steps.

(i) Observe that $\mathcal{J}_\alpha(\cdot, Q)$, in addition to being continuous, is also continuously differentiable. Indeed, we have

$$\begin{aligned} & \frac{\partial}{\partial P(x)} \mathcal{J}_\alpha(P, Q) \\ &= \frac{\alpha}{1-\alpha} \left[\frac{Q(x)^{\alpha-1}}{\sum_a P(a) Q(a)^{\alpha-1}} - \frac{P(x)^{\alpha-1}}{\sum_a P(a)^\alpha} \right]. \end{aligned} \quad (16)$$

Both denominators are bounded away from zero because for any $P \in \mathbb{L}$, we have $\max_x P(x) \geq 1/|\mathbb{X}|$, and therefore

$$\sum_a P(a) Q(a)^{\alpha-1} \geq \frac{1}{|\mathbb{X}|} \cdot \min_a Q(a)^{\alpha-1} > 0,$$

and

$$\sum_a P(a)^\alpha \geq \frac{1}{|\mathbb{X}|^\alpha} > 0.$$

Consequently, the partial derivative (16) exists everywhere on $\mathbb{R}_+^{|\mathbb{X}|}$, and is continuous because the terms involved are continuous. (The numerator of the second term in (16) is continuous because $\alpha > 1$).

(ii) Since the equality constraints in (13) and (14) arise from affine functions, and the inequality constraints in (15) arise from linear functions, we may apply [20, Prop. 3.3.7] to conclude that there exist Lagrange multipliers $(\lambda_i, i = 1, \dots, k)$, ν , and $(\mu(x), x \in \mathbb{X})$ associated with the constraints (13), (14), and (15), respectively, that satisfy:

$$\begin{aligned} & \frac{\alpha}{1-\alpha} \left[\frac{P^*(x)^{\alpha-1}}{\sum_a P^*(a)^\alpha} - \frac{Q(x)^{\alpha-1}}{\sum_a P^*(a) Q(a)^{\alpha-1}} \right] \\ &= \sum_{i=1}^k \lambda_i f_i(x) - \mu(x) + \nu \quad \forall x \end{aligned} \quad (17)$$

$$\mu(x) \geq 0 \quad \forall x \quad (18)$$

$$\mu(x) P^*(x) = 0 \quad \forall x. \quad (19)$$

In writing (17), we have substituted (16) for $\frac{\partial}{\partial P(x)} \mathcal{J}_\alpha(P, Q)$.

(iii) Multiplying (17) by $P^*(x)$, summing over all $x \in \mathbb{X}$, using $P^* \in \mathbb{L}$, and (19), we see that $\nu = 0$.

(iv) If $P^*(x) > 0$, we must have $\mu(x) = 0$ from (19), and its substitution in (17) yields, for all such x ,

$$\begin{aligned} & \frac{P^*(x)^{\alpha-1}}{\sum_a P^*(a)^\alpha} \\ &= \frac{Q(x)^{\alpha-1}}{\sum_a P^*(a) Q(a)^{\alpha-1}} + \frac{1-\alpha}{\alpha} \sum_{i=1}^k \lambda_i f_i(x). \end{aligned} \quad (20)$$

If $P^*(x) = 0$, (17) implies that

$$\begin{aligned} & \frac{Q(x)^{\alpha-1}}{\sum_a P^*(a) Q(a)^{\alpha-1}} + \frac{1-\alpha}{\alpha} \sum_{i=1}^k \lambda_i f_i(x) \\ &= \frac{(1-\alpha)}{\alpha} \mu(x) \\ &\leq 0, \end{aligned} \quad (21)$$

where the last inequality holds because of (18) and $\alpha > 1$. Therefore, (20) and (21) may be combined as

$$\begin{aligned} Z^{\alpha-1} P^*(x)^{\alpha-1} &= \left[Q(x)^{\alpha-1} + (1-\alpha) \sum_{i=1}^k \theta_i^* f_i(x) \right]_+ \\ &\quad \forall x \in \mathbb{X}, \end{aligned}$$

for obvious choices of Z and θ_i^* . This verifies (10) and completes the proof of (b).

(c) Using (10), for any $P \in \mathbb{L}$, we have

$$\begin{aligned} & \sum_x P(x) P^*(x)^{\alpha-1} \\ &\geq Z^{1-\alpha} \sum_x P(x) \left[Q(x)^{\alpha-1} + (1-\alpha) \sum_{i=1}^k \theta_i^* f_i(x) \right] \\ &= Z^{1-\alpha} \sum_x P(x) Q(x)^{\alpha-1}, \end{aligned} \quad (22)$$

where the first inequality follows because the terms on the right-hand side corresponding to x with $P^*(x) = 0$ are nonpositive, and the last equality follows because $P \in \mathbb{L}$. Similarly,

$$\sum_x P^*(x)^\alpha = Z^{1-\alpha} \sum_x P^*(x) Q(x)^{\alpha-1}.$$

Combining the above and (22) we get

$$\sum_x P(x) Q(x)^{\alpha-1} \leq \frac{\sum_x P^*(x) Q(x)^{\alpha-1}}{\sum_x P^*(x)^\alpha} \sum_x P(x) P^*(x)^{\alpha-1}, \quad (23)$$

which, using (1), is equivalent to (11). This completes the proof of (c).

(d) Same as proof of Theorem 2-(d).

(e) If $\text{Supp}(P^*) = \text{Supp}(\mathbb{L})$, then we have equality in (22), hence equality in (23), and hence equality in (11). This concludes the proof of (e) and the theorem. ■

As in the $\alpha < 1$ case, one has a converse.

Theorem 5: Let $\alpha > 1$. Let $P^* \in \mathbb{L}$ be a probability distribution of the form (10). Then P^* satisfies (11) for every $P \in \mathbb{L}$, and P^* is the \mathcal{I}_α -projection Q on \mathbb{L} . \square

Proof: The proof of Theorem 4-(c) shows that (11) holds. Now, for any $P \in \mathbb{L}$, we have

$$\begin{aligned}\mathcal{I}_\alpha(P, Q) &\geq \mathcal{I}_\alpha(P, P^*) + \mathcal{I}_\alpha(P^*, Q) \\ &\geq \mathcal{I}_\alpha(P^*, Q),\end{aligned}$$

which implies that P^* is the \mathcal{I}_α -projection Q on \mathbb{L} . \blacksquare

III. CONCLUDING REMARKS

Motivated by the maximum entropy principle, we studied minimization of a parametric extension of relative entropy, where the minimization was with respect to the first argument. We recognize that the minimizer on a set determined by linear statistical constraints belongs to the α -power-law family. This is analogous to the fact that the minimizer of relative entropy ($\alpha = 1$) belongs to the exponential family. A complementary minimization problem, where the minimization is with respect to the second argument of \mathcal{I}_α , is relevant in robust statistics ($\alpha > 1$) and in constrained compression settings ($\alpha < 1$). This problem and the connection between the two minimization problems will be the subject matter of our forthcoming work.

ACKNOWLEDGMENTS

M. Ashok Kumar was supported by a Council for Scientific and Industrial Research (CSIR) fellowship and by the Department of Science and Technology. R. Sundaresan was supported in part by the University Grants Commission by Grant Part (2B) UGC-CAS-(Ph.IV) and in part by the Department of Science and Technology.

REFERENCES

- [1] E. T. Jaynes, *Papers on Probability, Statistics and Statistical Physics*, R. D. Rosenkrantz, Ed. P.O. Box 17,3300 AA Dordrecht, The Netherlands.: Kluwer Academic Publishers, 1982.
- [2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2006.
- [3] J. M. V. Campenhout and T. M. Cover, "Maximum entropy and conditional probability," *Information Theory, IEEE Transactions on*, vol. 27, no. 4, pp. 483–489, July 1981.
- [4] I. Csiszár, "Sanov property, generalized I -projection, and a conditional limit theorem," *Ann. Prob.*, vol. 12, no. 3, pp. 768–793, 1984.
- [5] —, " I -divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, pp. 146–158, 1975.
- [6] I. Csiszár and F. Matúš, "Information projections revisited," *Information Theory, IEEE Transactions on*, vol. 49, no. 6, pp. 1474–1490, June 2003.
- [7] L. L. Campbell, "A coding theorem and Rényi's entropy," *Information and Control*, vol. 8, pp. 423–429, 1965.
- [8] A. C. Blumer and R. J. McEliece, "The Rényi redundancy of generalized Huffman codes," *Information Theory, IEEE Transactions on*, vol. 34, no. 5, pp. 1242–1249, September 1988.
- [9] R. Sundaresan, "Guessing under source uncertainty," *Information Theory, IEEE Transactions on*, vol. 53, no. 1, pp. 269–287, January 2007.
- [10] E. Arikan, "An inequality on guessing and its application to sequential decoding," *Information Theory, IEEE Transactions on*, vol. 42, no. 1, pp. 99–105, January 1996.
- [11] M. K. Hanawal and R. Sundaresan, "Guessing revisited: A large deviations approach," *Information Theory, IEEE Transactions on*, vol. 57, no. 1, pp. 70–78, January 2011.
- [12] R. Sundaresan, "A measure of discrimination and its geometric properties," in *Proc. of the 2002 IEEE International Symposium on Information Theory*, Lausanne, Switzerland, June 2002, p. 264.
- [13] C. Bunte and A. Lapidoth, "Codes for tasks and Rényi entropy," *Information Theory, IEEE Transactions on*, vol. 60, no. 9, pp. 5065–5076, September 2014.
- [14] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis*, vol. 99, pp. 2053–2081, 2008.
- [15] O. T. Johnson and C. Vignat, "Some results concerning maximum Rényi entropy distributions," *Annales de l'Institut Henri Poincaré (B)*, vol. 43, no. 3, pp. 339–351, May-June 2007.
- [16] E. Lutwak, D. Yang, and G. Zhang, "Cramer-Rao and moment-entropy inequalities for Rényi entropy and generalized Fisher information," *Information Theory, IEEE Transactions on*, vol. 51, no. 1, pp. 473–478, January 2005.
- [17] J. Costa, A. Hero, and C. Vignat, "On solutions to multivariate maximum-entropy problems," in *EMMCVPR 2003, Lisbon, Portugal*, ser. Lecture Notes in Computer Science, A. Rangarajan, M. Figueiredo, and J. Zerubia, Eds., vol. 2683. Berlin, Germany: Springer-Verlag, July 2003, pp. 211–228.
- [18] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, vol. 52, no. 1-2, pp. 479–487, 1988.
- [19] I. Csiszár and P. Shields, *Information Theory and Statistics: A Tutorial*, ser. Foundations and Trends in Communications and Information Theory. Hanover, USA: Now Publishers Inc, 2004, vol. 1, no. 4.
- [20] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA: Athena Scientific, 2003.